

Design And Development of an NLP-Based Chatbot For Information Services

Sean Andrianto¹, Ferri Krisdiantoro², Putri Rizqiyah³, Kusnadi⁴

^{1,2,4} Informatics Engineering, Information Technology Department, Catur Insan Cendekia University

³ Information System, Telkom University Jakarta

Correspondence: seanandrianto02@gmail.com, ferri.krisdiantoro.study@gmail.com, puputrizqiyah24@gmail.com
kusnadi@cic.ac.id,

Article Info

Article history:

Received Jun 12th, 2026

Revised Aug 20th, 2026

Accepted Aug 26th, 2026

Keyword:

Chatbot; Natural Language Processing; IndoBERT; Cosine Similarity; Information Service.

ABSTRACT

Pondok Pesantren Modern AL-Ikhlas Putri faces difficulty in providing information services that are accessible at any time, since existing information channels rely on staff who can only be reached during limited working hours, making it hard for students, teachers, prospective students' parents, and the general public to obtain quick answers outside those hours. This study aims to design and build an NLP-based information service chatbot that functions as a round-the-clock virtual assistant for answering questions about the pesantren's activities, curriculum, facilities, and general information. The system was developed using a three-tier architecture consisting of a user interface (React.js), a Backend-for-Frontend server (Node.js), and an AI/ML service (Flask). A Natural Language Processing approach was applied by leveraging the pre-trained IndoBERT model to transform user questions into vector embeddings, and Cosine Similarity was then used to measure the semantic similarity between a user's query and an FAQ-formatted knowledge base. Black box testing showed that the chatbot was able to return accurate and relevant answers at any time, including for varied sentence structures and questions containing typographical errors, confirming that the semantic retrieval mechanism functioned effectively. The resulting chatbot provides an innovative digital solution that makes information about the pesantren accessible around the clock, supporting digital transformation in pesantren-based education.



© 2026 Sean Andrianto, Ferri Krisdiantoro, Putri Rizqiyah, Kusnadi. Published by Permata Harmoni Abadi. This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>)

INTRODUCTION

Information technology is developing rapidly in this digital era, bringing significant changes to various sectors, including education. Digitalization does not only involve improving infrastructure, networks, and software, but also opens opportunities for educational institutions to provide services and information that are accessible at any time, not only during conventional working hours (Maharani & Wardhana, 2024). One form of this transformation is the application of artificial intelligence (AI)-based technology to support round-the-clock access to information.

Pondok Pesantren Modern Al-Ikhlas Putri is an educational institution that combines religious knowledge with general knowledge, and consequently receives a large volume of information requests from students, teachers, prospective students' parents, and the general public regarding its activities, curriculum, facilities, and admission procedures. In practice, most of these questions can currently only be answered directly by pesantren staff, who are available solely during limited office hours; questions submitted outside these hours, for example at night, on holidays, or when staff are occupied with other duties, often go unanswered or must wait until the next working day. The pesantren's website has so far provided static content such as blog posts and a glossary of terms, but this content cannot respond interactively to a user's specific question, so it does not resolve the underlying problem of the unavailability of an information service around the clock. A similar limitation was reported by Hikmah et al. (2022), who found that a campus information service unable to operate 24 hours a day created a significant obstacle for users seeking quick answers, and addressed the problem by developing a chatbot as a virtual assistant. Building on this insight, a technology-based solution is needed that can provide continuous, interactive access to information without depending on staff availability.

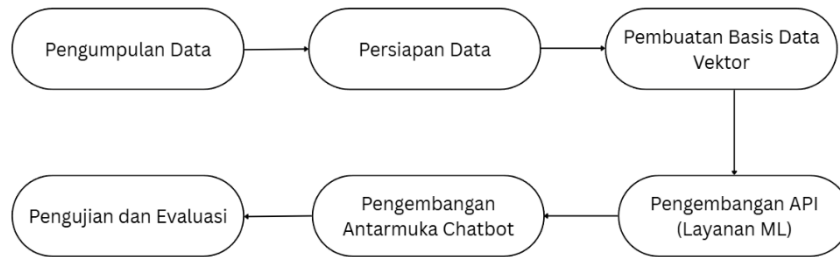
One innovation that can be applied to address this problem is the development of a chatbot based on Natural Language Processing (NLP) that functions as an automated information-service assistant, available at any time, to answer various questions regarding the pesantren's activities, curriculum, facilities, and general information. This study utilizes the IndoBERT model, chosen for its superior ability to understand the context of the Indonesian language (Pramana & Romadhony, 2024), to transform question text into vector representations (embeddings) that are then measured for similarity using the Cosine Similarity method to find the most appropriate answer from the pesantren's knowledge base. The system was built using a modern architecture that separates Flask as the backend for NLP model inference, Node.js as the Backend-for-Frontend (BFF), and React.js as an interactive and responsive user interface (Khan et al., 2021), in line with the pesantren's efforts to adopt technology to support the digital transformation of pesantren-based education (Umar et al., 2023).

Previous studies have shown various approaches to developing information-service chatbots. Fahadra (2023) developed a final-project information-service chatbot using the Long Short-Term Memory (LSTM) method and achieved an accuracy of 94.41%; however, LSTM-based approaches have limitations in understanding complex sentence context compared with more modern Transformer-based architectures. Hikmah et al. (2022) implemented a chatbot based on Term Frequency-Inverse Document Frequency (TF-IDF) and Cosine Similarity to overcome the same 24-hour service limitation at a university, achieving an accuracy of 85.7%, but this approach focuses only on keyword matching, making it prone to failure when a user's question does not contain keywords identical to those in the database. Ibadurrahman et al. (2025) applied fine-tuning of the BERT model for intent classification in a dental-clinic chatbot with an accuracy of 86%, while Wintoro et al. (2022) reported an accuracy of 99% for an LSTM-based academic information chatbot. In the broader field of natural language processing, Rizqiyah (2024) showed that data-augmentation techniques can improve the performance of a Transformer-based classification model (XLNet) up to an accuracy of 74.37% on a sentiment-analysis task.

Unlike these previous studies, this research uses IndoBERT, a Transformer model specifically trained on a large-scale Indonesian-language corpus, making it superior in understanding nuances and terms specific to the pesantren domain compared with general-purpose LSTM or BERT models. This study also combines this approach with a three-tier architecture that separates the user interface, intermediary server, and AI model service, and utilizes an embedding-extraction strategy without fine-tuning, making it more efficient to apply in domains with limited data. Based on the above, this study aims to design and build an NLP-based information service chatbot using the IndoBERT model and the Cosine Similarity method, integrate it into a reliable and scalable three-tier architecture, and evaluate the system's functionality through black box testing. The contribution of this study is to provide an innovative digital solution that makes information about the pesantren accessible to students, teachers, and the general public at any time, thereby overcoming the limitation of an information service that was previously available only during working hours, while also supporting digital transformation in pesantren-based educational environments.

RESEARCH METHODS

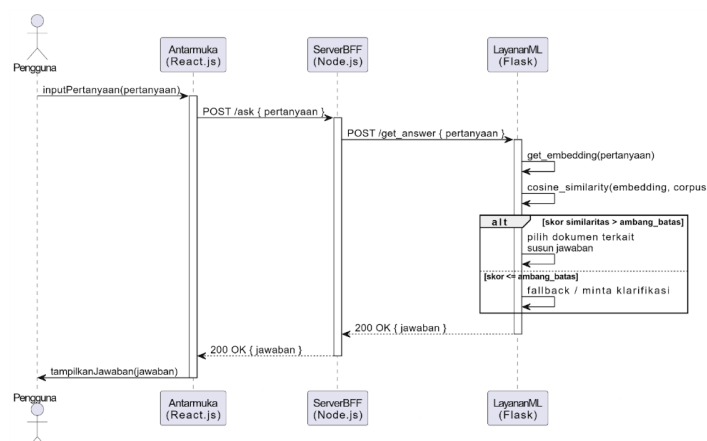
This study is applied research aimed at producing a software product in the form of an NLP-based information service chatbot, with a case study at Pondok Pesantren Modern Al-Ikhlas Putri, located in Ciomas Village, Ciawigebang District, Kuningan Regency, West Java. The object of the research is the need for round-the-clock information access among students, teachers, and the general public regarding the pesantren's activities, curriculum, facilities, admission procedures, and general terminology. The research workflow is divided into two main stages: an offline development stage focused on data and model preparation, and an online system development stage focused on building the application architecture and real-time inference, as illustrated in Figure 1.



Picture 1 Research Methodology Flow Diagram

The chatbot's knowledge data was obtained through direct interviews with the management of Pondok Pesantren Modern Al-Ikhlâs Putri to collect frequently asked questions (FAQs) that users typically need answered at any time, covering activities, curriculum, facilities, admission procedures, specific terminology, and other general information about the pesantren. The raw data from the interviews was then structured into a question-and-answer format and stored as a knowledge base (faq.json). In the data-preparation stage, each question in the database was converted into a vector representation (embedding) using the pre-trained IndoBERT model (indobenchmark/indobert-base-p1) without a fine-tuning process, and the result was stored in a binary file (faq_embeddings.npy) so that the answer-search process at runtime would not need to recompute embeddings for the entire database every time a new question was received.

The system was designed using a three-tier architecture consisting of three components. First, the Frontend layer was built using React.js and is responsible for the user interface in the form of an interactive chatbot modal that can be activated via an icon on the pesantren's website. Second, the Backend-for-Frontend (BFF) layer was built using Node.js and functions as an intermediary server that forwards requests from the Frontend to the AI service while also preventing direct client access to the model, thereby improving system security. Third, the Backend AI/ML layer was built using the Flask (Python) micro-framework, which exposes an endpoint (/get_answer) to receive question text, generate embeddings using the IndoBERT model, calculate the Cosine Similarity between the user's question embedding and all question embeddings in the database, and then return the answer with the highest similarity score if it exceeds a certain threshold. The interaction between these three layers is shown in Figure 2, which illustrates the flow of a request from the user until the answer is displayed back on the interface.



Picture 2 Sequence Diagram of Inter-Layer System Interaction

System evaluation was carried out using the Black Box Testing method, namely testing the system's functionality from the user's point of view without examining the internal code structure. The test scenarios were designed to cover relevant questions, questions with varied sentence structures,

questions irrelevant to the pesantren domain, empty input, and questions containing typographical errors (typos), in order to assess whether the semantic-similarity-based search mechanism could consistently find the correct answer, in line with the functional and non-functional requirements formulated during the requirements-analysis stage.

RESULTS AND DISCUSSION

The information service chatbot was successfully implemented in accordance with the three-tier architecture design described in the Research Methods section. The implementation environment used in this study is presented in Table 1, which lists the software versions for each system layer, ranging from the Python environment and model-processing libraries at the Backend AI/ML layer to the interface framework at the Frontend layer

Table 1 Software Implementation Environment

Software	Version
Python	3.12.3
React.js	18.2.0
Tailwind CSS	3.4.116
Node.js (LTS)	18.20.8
TensorFlow	2.19.0
Transformers (Hugging Face)	4.51.3
Flask	3.1.0

Functional testing was carried out using the Black Box Testing method with five scenarios representing input variations that the system might realistically receive under actual, round-the-clock use, as summarized in Table 2. The data in Table 2 show that the system was not tested using raw data; rather, each input was first processed through the embedding and Cosine Similarity computation stages before an answer was displayed to the user.

Table 2 Black Box Testing Results of the Information Service Chatbot

No	Test Scenario	Example User Input	Result
1	Relevant question	“Apa saja fasilitas yang tersedia?” (“What facilities are available?”)	As expected
2	Question with varied sentence structure	“Fasilitas yang ditawarkan pondok apa saja?” (“What facilities does the pesantren offer?”)	As expected
3	Irrelevant question	“Siapa nama presiden ke-8 Indonesia?” (“Who was the 8th president of Indonesia?”)	As expected (default message displayed)
4	Empty input	Send button pressed without typing anything	As expected (request not sent)
5	Question with a typographical error (typo)	“fsilita pondok yang tersedia?” (typo of “fasilitas pondok yang tersedia?”)	As expected

The test results in Table 2 show that the chatbot was able to find appropriate answers for relevant questions (scenario 1) as well as for questions with varied sentence structures (scenario 2) that were lexically different from the questions in the database but carried the same meaning. This demonstrates that the IndoBERT embedding and Cosine Similarity approach is able to capture semantic similarity rather than mere keyword matching, as in the TF-IDF approach reported by Hikmah et al.

(2022), who noted that TF-IDF-based systems tend to fail to find answers when a user's question does not contain keywords identical to those in the database.

In scenario 3, the system successfully recognized a question outside the pesantren domain and displayed a default message, indicating that the similarity-threshold mechanism functioned as intended, so the chatbot did not provide a misleading answer when there was no adequate match in the database. In scenario 4, empty-input validation on the interface side prevented an empty request from being sent to the server, indicating that the non-functional requirement related to usability had been met. A finding that specifically supports the advantage of the embedding-based approach is shown in scenario 5, in which the chatbot was still able to find the correct answer even though the user's question contained a typographical error. This capability is consistent with the characteristics of contextual vector representations in Transformer models such as IndoBERT (Vaswani et al., 2017; Wilie et al., 2020), which map sentence meaning into a high-dimensional vector space so that small variations on the text surface do not always significantly change the vector's position, unlike purely lexical-matching approaches, which are far more vulnerable to typographical errors.

Compared with previous studies that reported quantitative accuracy, such as an LSTM-based chatbot with an accuracy of 94.41% (Fahadra, 2023) and a BERT-based chatbot with an accuracy of 86% in the dental-clinic domain (Ibadurrahman et al., 2025), this study did not include a quantitative accuracy measurement, since the testing performed was qualitative in nature through a limited set of Black Box Testing scenarios. This is a limitation of the study that should be addressed in future work through testing with a larger sample of questions and quantitative evaluation metrics such as accuracy, precision, recall, or F1-score, so that the system's performance can be compared more rigorously with similar studies. Nevertheless, from an architectural standpoint, separating the system into three independent layers (React.js, Node.js, and Flask) proved to simplify management and further development, since updating the knowledge base only requires updating the `faq.json` file and re-running the embedding-generation process, without needing to modify the interface or intermediary-server components, in line with the separation-of-concerns principle recommended in modern service architectures (Cahn, 2017). More importantly, because the chatbot runs as an automated service that does not depend on staff availability, it is able to answer user questions at any time, including outside office hours, on holidays, or at night, thereby directly addressing the round-the-clock information access problem that motivated this study.

CONCLUSION

This study successfully designed and built an NLP-based information service chatbot for Pondok Pesantren Modern Al-Ikhlash Putri, using the IndoBERT model to generate question embeddings and the Cosine Similarity method to measure semantic similarity, implemented on a three-tier architecture consisting of React.js as the user interface, Node.js as the Backend-for-Frontend, and Flask as the AI model service. The Black Box Testing results show that the system was able to consistently find relevant answers at any time, both for questions matching the database and for questions with varied sentence structures or typographical errors, while questions outside the pesantren domain were handled appropriately through the similarity-threshold mechanism. By operating without dependence on staff availability, the resulting system directly overcomes the pesantren's original problem of not having an information service accessible outside limited office hours, providing an innovative digital solution that makes information quickly and continuously accessible to students, teachers, and the general public while supporting digital transformation in pesantren-based educational environments, although further evaluation using quantitative accuracy metrics on a larger sample is still needed to measure the system's performance more comprehensively.

REFERENCES

Boykis, V. (2023). What are embeddings? Zenodo. <https://doi.org/10.5281/zenodo.8015029>

- Cahn, J. (2017). *Chatbot: Architecture, design, & development* [Master's thesis, University of Pennsylvania].
- El Naqa, I., & Murphy, M. J. (2015). What is machine learning? In *Machine learning in radiation oncology: Theory and applications* (pp. 3–11). Springer.
- Fahadra, L. (2023). *Sistem chatbot layanan informasi tugas akhir menggunakan metode Long Short-Term Memory (LSTM)* [Skripsi, Universitas Islam Sultan Agung].
- Hikmah, N., Ariyanti, D., & Pratama, F. A. (2022). Implementasi chatbot sebagai virtual assistant di Universitas Panca Marga Probolinggo menggunakan metode TF-IDF. *JTIM: Jurnal Teknologi Informasi dan Multimedia*, 4(2), 133–148.
- Ibadurrahman, I., Yusuf, F., & Budianto, H. (2025). Implementasi chatbot berbasis Natural Language Processing (NLP) menggunakan model BERT untuk meningkatkan pelayanan klinik gigi. *Rabit: Jurnal Teknologi dan Sistem Informasi Univrab*, 10(2), 325–334. <https://doi.org/10.36341/rabit.v10i2.6061>
- Khan, F. S., Al Mushabbir, M., Irbaz, M. S., & Al Nasim, M. D. A. (2021). End-to-end natural language understanding pipeline for Bangla conversational agents. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 205–210). IEEE.
- Maharani, D., & Wardhana, K. E. (2024). Digitalisasi sistem administrasi sekolah dengan pembuatan website di SD Negeri Loa Kulu. *Novara: Nusantara Education and Innovation Journal*, 1(3), 133–141.
- Manning, C. D. (2008). *Introduction to information retrieval*. Syngress Publishing.
- Mulyatun, S., Utama, H., & Mustopa, A. (2021). Pendekatan Natural Language Processing pada aplikasi chatbot sebagai alat bantu customer service. *Journal of Information System Management (JOISM)*, 2(2), 12–17.
- Pramana, R. A., & Romadhony, A. (2024). Identifikasi similar question dengan IndoBERT (Studi kasus dataset QAS Covid-19). *LOGIC: Jurnal Penelitian Informatika*, 2(1), 12–17.
- Rizqiyah, P. (2024). *Analisis sentimen vaksin Covid-19 di Indonesia menggunakan model bahasa XLMR dan teknik augmentasi data* [Tesis, Universitas Indonesia].
- Umar, T. M., Chaerowati, D. L., & Drajat, M. S. (2023). Digital literacy of santri through Islamic boarding school culture. *KnE Social Sciences*, 380–392.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. *arXiv preprint arXiv:2009.05387*.
- Wintoro, P. B., Hermawan, H., Muda, M. A., & Mulyani, Y. (2022). Implementasi Long Short-Term Memory pada chatbot informasi akademik Teknik Informatika Unila. *Expert: Jurnal Manajemen Sistem Informasi dan Teknologi*, 12(1), 68.